Deep learning based Feature Selection Study

Sanghyuck Lee

2021. 08 10.

Outline

- Knockoffs Pruning for Segmentation Model
- 중견 연구 과제
- 갑상선 연구 과제
- 산학 협력 과제

Paper Survey

1) Search Keywords: Knockoff, Knockoff filter, Deep learning, Neural Network

Paper Title: Deep knockoffs [1]

This article introduces a machine for **sampling approximate model-X knockoffs** for arbitrary and unspecified data distributions using **deep generative models**. The main idea is to iteratively refine a knockoff sampling mechanism until a criterion measuring the validity of the produced knockoffs is optimized; this criterion is inspired by the popular maximum mean discrepancy in machine learning and can be thought of as measuring the distance to pairwise exchangeability between original and knockoff features. By building upon the existing model-X framework, we thus obtain a flexible and model-free statistical tool to perform controlled variable selection. Extensive numerical experiments and quantitative tests confirm the generality, effectiveness, and power of our deep knockoff machines. Finally, we apply this new method to a real study of mutations linked to changes in drug resistance in the human immunodeficiency virus. Supplementary materials for this article are available online.

Paper Survey

Paper Title: Deep direct likelihood knockoffs [2]

Predictive modeling often uses black box machine learning methods, such as deep neural networks, to achieve state-of-the-art performance. In scientific domains, the scientist often wishes to discover which features are actually important for making the predictions. These discoveries may lead to costly follow-up experiments and as such it is important that the error rate on discoveries is not too high. Model-X knockoffs [2] enable important features to be discovered with control of the false discovery rate (fdr). **However, knockoffs require rich generative models** capable of accurately modeling the knockoff features while ensuring they obey the so-called "swap" property. We develop Deep Direct Likelihood Knockoffs (ddlk), which directly minimizes the KL divergence implied by the knockoff swap property. ddlk consists of two stages: it first maximizes the explicit likelihood of the features, then minimizes the KL divergence between the joint distribution of features and knockoffs and any swap between them. **To ensure that the generated knockoffs are valid** under any possible swap, ddlk uses the Gumbel-Softmax trick to optimize the knockoff generator under the worst-case swap. We find ddlk has higher power than baselines while controlling the false discovery rate on a variety of synthetic and real benchmarks including a task involving a large dataset from one of the epicenters of COVID-19.

Paper Survey

Paper Title: Knockoff-Inspired Feature Selection via Generative Models [3]

We **propose a feature selection algorithm** for supervised learning inspired by the recently introduced knockoff framework for variable selection in statistical regression. While variable selection in statistics aims to distinguish between true and false predictors, feature selection in machine learning aims to reduce the dimensionality of the data while preserving the performance of the learning method. The knockoff framework has attracted significant interest due to its strong control of false discoveries while preserving predictive power. In contrast to the original approach and later variants that assume a given probabilistic model for the variables, **our proposed approach relies on data-driven generative models** that learn mappings from data space to a parametric space that characterizes the probability distribution of the data. Our approach requires only the availability of mappings from data space to a distribution in parametric space and from parametric space to a distribution in data space; thus, it can be integrated with multiple popular generative models from machine learning. We provide example knockoff designs using a variational autoencoder and a Gaussian process latent variable model. We also propose a knockoff score metric for a softmax classifier that accounts for the contribution of each feature and its knockoff during supervised learning. Experimental results with **multiple benchmark datasets for feature selection** showcase the advantages of our knockoff designs and the knockoff framework with respect to existing approaches.

Paper Survey

Paper Title: Distribution-dependent feature selection for deep neural networks [4]

While deep neural networks (DNNs) have achieved impressive performance on a wide variety of tasks, the black-box nature hinders their applicability to high-risk, decision-making fields. In such fields, besides accurate prediction, it is also desired to provide interpretable insights into DNNs, e.g., screening important features based on their contributions to predictive accuracy. To improve the interpretability of DNNs, this paper originally **proposes a new feature selection algorithm for DNNs** by integrating the knockoff technique and the distribution information of irrelevant features. With the help of knockoff features and central limit theorem, we state that the irrelevant feature's statistic follows a known Gaussian distribution under mild conditions. This information is applied in hypothesis testing to **discover key features associated with the DNNs**. Empirical evaluations on simulated data demonstrate that the proposed method can select more true informative features with **higher F1 scores**. Meanwhile, the Friedman test and the post-hoc Nemenyi test are employed to validate the superiority of the proposed method. Then we apply our method to Coronal Mass Ejections (CME) data and uncover the key features which contribute to the DNN-based CME arrival time.

Paper Survey

Paper Title: Supervised feature selection through Deep Neural Networks with pairwise connected structure [5]

Feature selection is an important data preprocessing strategy, has been proven empirically that it contributes to reducing the dimensionality of feature and enhancing the performance of learning algorithms in practice. Typical sparse learning-based models select the features by removing ones that the feature scores are zero. However, linear models puzzle to build the non-linear relations between features and responses. The Deep Neural Network (DNN) has a strong capability to mode the non-linear relations and has been employed to select features. In this paper, **we introduce a novel deep Neural network-based Feature Selection (NeuralFS) method to identify features.** The new model comprises of a **fully-connection network, a decision network**, and connect them through a pairwise connected structure. In NeuralFS, the **fully-connected network** is the crucial structure in NeuralFS that **transforms the features into their corresponding scores**, and the **decision network** is the final structure that **performs classification or regression.** The **pairwise connected can be regarded as a "bridge"** to connect the two networks, and its weights are fixed as the normalized input as well as it is un-trainable during model training. After optimizing, the feature scores can be obtained by calculating the output of the fully-connected network. NeuralFS takes advantage of the deep network to model the non-linearity, and also make features scores sparse without the sparse regularization technology. We apply the proposed method to both synthetic datasets and benchmark datasets to prove its effectiveness.

Paper Survey

Paper Title: Deep-gKnock: nonlinear group-feature selection with deep neural network [6]

Feature selection is central to contemporary high-dimensional data analysis. Group structure among features arises naturally in various scientific problems. Many methods have been proposed to incorporate the group structure information into feature selection. However, these methods are normally restricted to a linear regression setting. To relax the linear constraint, **we design a new Deep Neural Network (DNN) architecture** and integrating it with the recently proposed knockoff technique to **perform nonlinear group-feature selection** with controlled group-wise False Discovery Rate (gFDR). Experimental results on high-dimensional synthetic data demonstrate that our method **achieves the highest power and accurate gFDR control** compared with state-of-the-art methods. The performance of Deep-gKnock is especially superior in the following five situations: (1) nonlinearity relationship; (2) dimension greater than sample size ; (3) high between-group correlation; (4) high within-group correlation; (5) large number of associated groups. And Deep-gKnock is also demonstrated to be robust to the misspecification of the feature distribution and the change of network architecture. Moreover, Deep-gKnock achieves scientifically meaningful group-feature selection results for cutting-edge real world datasets.

Paper Survey

Paper Title: An Effective Pulmonary Nodule Detection In Ctand X-Ray Images Using Enhanced Inception-Residual Convolutional Neural Network [7]

Purpose: **Lung Nodule** or Pulmonary Nodule is a space in white taking after we identify the cotton or mists in clinical pictures examined on thorax. **Early identification of nodules of lung** in thoracic **Computed Tomography (CT)** examines that it is critical for the effective conclusion and treatment of lung disease. Prior, the grouping is finished through the Entropy Weighted Residual Convolution Neural Network (EWRCNN) for recognition of lung nodule. This identification plan can maintain a strategic distance from competitor extraction and be less reliant on scale. In any case, division of suspected aspiratory nodules requires further innovative work which can expand the recognition affectability of detection systems.

Paper Survey

Paper Title: An Effective Pulmonary Nodule Detection In Ctand X-Ray Images Using Enhanced Inception-Residual Convolutional Neural Network [7]

Contribution: To accomplish this objective, Enhanced Inception-Residual Convolutional Neural Network (EIRCNN) is proposed which has three stages, for example, pre-preparing, segmentation and final classification of pulmonary nodule. At first, to improve the Ct images local information, Normalized Gamma-Corrected Contrast-Limited Adaptive Histogram Equalization (NGCCLAHE)- Discrete Wavelet Transform (DWT) is recommended that consolidates the NGCCLAHE with DWT. Second, **intelligent W-Net** (iW-Net) is a profound learning model that takes into consideration both programmed and **intuitive division of lung nodules** in processed tomography pictures following that the feature extraction is directed utilizing the **Fuzzy Continuous Wavelet Transform** (FCWT) and **Gray Level Feature Extraction** (GLCM). And then, **feature selection** deducts the **filter-based approach** to deal with production of the ideal list of capabilities and to limit the immateriality of the features that gives the output for the lung nodules classification in CT and X-ray images. The arrangement is finished utilizing EIRCNN.

Results: The strategy EIRCNN is profoundly powerful for decreasing the false positive rates, on 888 scans of the freely accessible LIDC-IDRI dataset, when contrasted with the existing techniques, for example, EWRCNN, Faster RCNN and RCNN.

Paper Survey

Paper Title: Interpreting black box models via hypothesis testing [8]

In science and medicine, model interpretations may be reported as discoveries of natural phenomena or used to guide patient treatments. In such high-stakes tasks, false discoveries may lead investigators astray. These applications would therefore benefit from control over the finite-sample error rate of interpretations. We **reframe black box model interpretability as a multiple hypothesis testing problem. The task is to discover "important" features** by testing whether the model prediction is significantly different from what would be expected if the features were replaced with uninformative counterfactuals. **We propose two testing methods**: one that provably **controls the false discovery rate** but which is **not yet feasible for large-scale applications**, and an **approximate testing method which can be applied to real-world data sets.** In simulation, both tests have high power relative to existing interpretability methods. When applied to state-of-the-art **vision and language models**, the framework selects features that intuitively explain model predictions. The resulting explanations have the additional advantage that they are themselves easy to interpret.

Paper Survey

Paper Title: KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks [9]

Feature selection is a pervasive problem. The discovery of relevant features can be as important for performing a particular task (such as to avoid overfitting in prediction) as it can be for understanding the underlying processes governing the true label (such as discovering relevant genetic factors for a disease). Machine learning driven feature selection can enable discovery from large, high-dimensional, nonlinear observational datasets by creating a subset of features for experts to focus on. In order to use expert time most efficiently, we need a principled methodology capable of controlling the False Discovery Rate. In this work, **we build on the promising Knockoff framework by developing a flexible knockoff generation model.** We adapt the Generative Adversarial Networks framework to allow us to generate knockoffs with no assumptions on the feature distribution. Our model consists of 4 networks, a generator, a discriminator, a stability network and a power network. We demonstrate the capability of our model to perform feature selection, showing that it performs as well as the originally proposed knockoff generation model in the Gaussian setting and that it outperforms the original model in non-Gaussian settings, including on a real-world dataset.

Paper Survey

Paper Title: Variable selection with false discovery rate control in deep neural networks [10]

Deep neural networks (DNNs) are famous for their high prediction accuracy, but they are also known for their black-box nature and poor interpretability. We consider the problem of variable selection, that is, selecting the input variables that have significant predictive power on the output, in DNNs. **We propose a backward elimination procedure called SurvNet, which is based on a new measure of variable importance that applies to a wide variety of networks.** More importantly, SurvNet is **able to estimate and control the false discovery rate of selected variables**, while no existing methods provide such a quality control. Further, SurvNet adaptively determines how many variables to eliminate at each step in order to maximize the selection efficiency. To study its validity, SurvNet is applied to **image data** and gene expression data, as well as various simulation datasets.

Paper Survey

Paper Title: Causal inference using deep-learning variable selection identifies and incorporates direct and indirect causalities in complex biological systems [11]

In complex diseases, **causal structure learning across biological variables** is critical to identify modifiable triggers or potential therapeutic agents. A limitation of existing causal learning methods is that they cannot identify indirect causal relations, those that would interact through latent mediating variables. **We developed the first computational method that identifies both direct and indirect causalities, causal inference using deep-learning variable-selection (causalDeepVASE).** To accurately identify indirect causalities and incorporate them with direct causalities, causalDeepVASE develops a deep neural network approach and extends a flexible causal inference method. In simulated and biological data of various contexts, causalDeepVASE outperforms existing methods in identifying expected or validated causal relations. Further, causalDeepVASE facilitates a systematic understanding of complex diseases. For example, causalDeepVASE uniquely identified a possible causal relation between IFNγ and creatinine suggested in a polymicrobial sepsis model. In future biomedical studies, causalDeepVASE can facilitate the identification of driver genes and therapeutic agents.

Paper Survey

Paper Title: An embedded gene selection method using knockoffs optimizing neural network [12]

Background

Gene selection refers to **find a small subset of discriminant genes** from the gene expression profiles. How to select genes that affect specific phenotypic traits effectively is an important research work in the field of biology. The neural network has better fitting ability when dealing with nonlinear data, and it can capture features automatically and flexibly. **In this work, we propose an embedded gene selection method using neural network. The important genes can be obtained by calculating the weight coefficient after the training is completed.** In order to solve the problem of black box of neural network and further make the training results interpretable in neural network, we use the idea of knockoffs to construct the knockoff feature genes of the original feature genes. This method not only make each feature gene to compete with each other, but also make each feature gene compete with its knockoff feature gene. This approach can help to select the key genes that affect the decision-making of neural networks.

Paper Survey

Paper Title: An embedded gene selection method using knockoffs optimizing neural network [12]

Results

We use maize carotenoids, tocopherol methyltransferase, raffinose family oligosaccharides and human breast cancer dataset to do verification and analysis.

Conclusions

The experiment results demonstrate that the knockoffs optimizing neural network method has better detection effect than the other existing algorithms, and specially for processing the nonlinear gene expression and phenotype data.

Paper Survey

Paper Title: Auto-encoding knockoff generator for FDR controlled variable selection [13]

A new statistical procedure (Model-X Candes et al.(2018)) has provided a way to **identify important factors using any supervised learning method controlling for FDR.** This line of research has shown great potential to expand the horizon of machine learning methods beyond the task of prediction, to serve the broader needs in scientific researches for interpretable findings. **However, the lack of a practical and flexible method to generate knockoffs** remains the major obstacle for wide application of Model-X procedure. This paper fills in the gap by **proposing a model-free knockoff generator** which approximates the correlation structure between features through latent variable representation. We demonstrate our proposed method can achieve FDR control and better power than two existing methods in various simulated settings and a real data example for finding mutations associated with drug resistance in HIV-1 patients.

Paper Survey

Paper Title: INVASE: Instance-wise variable selection using neural networks [14]

The advent of big data brings with it data with more and more dimensions and thus a growing need to be able to efficiently select which features to use for a variety of problems. While global feature selection has been a well-studied problem for quite some time, only recently has the paradigm of instance-wise feature selection been developed. In this paper, **we propose a new instance-wise feature selection method, which we term INVASE.** INVASE consists of 3 neural networks, **a selector network, a predictor network and a baseline network** which are used to **train the selector network using the actor-critic methodology**. Using this methodology, INVASE is capable of **flexibly discovering feature subsets of a different size for each instance**, which is a key limitation of existing state-of-the-art methods. We demonstrate through a mixture of synthetic and real data experiments that INVASE significantly outperforms state-of-the-art benchmarks.

Paper Survey

Paper Title: Scop: Scientific control for reliable neural network pruning [15]

This paper proposes a reliable **neural network pruning** algorithm by setting up a scientific control. Existing pruning methods have developed various hypotheses to approximate the importance of filters to the network and then execute filter pruning accordingly. To increase the reliability of the results, we prefer to have a more rigorous research design by including a scientific control group as an essential part to minimize the effect of all factors except the association between the filter and expected network output. Acting as a control group, **knockoff feature is generated to mimic the feature map produced** by the network filter, but they are conditionally independent of the example label given the real feature map. We theoretically suggest that the knockoff condition can be approximately preserved given the information propagation of network layers. Besides the real feature map on an intermediate layer, the corresponding knockoff feature is brought in as another auxiliary input signal for the subsequent layers. **Redundant filters can be discovered** in the adversarial process of different features. Through experiments, we demonstrate the superiority of the proposed algorithm over state-of-the-art methods. For example, our method can **reduce** 57.8% **parameters** and 60.2% **FLOPs** of ResNet-101 with only 0.01% top-1 accuracy loss on ImageNet.

Paper Survey

Paper Title: Robust Decoy-enhanced Saliency Maps [16]

Saliency methods can make deep neural network predictions more interpretable by identifying a set of critical features in an input sample, such as pixels that contribute most strongly to a prediction made by an image classifier. Unfortunately, recent evidence suggests that many saliency methods poorly perform, especially in situations where gradients are saturated, inputs contain adversarial perturbations, or predictions rely upon interfeature dependence. To address these issues, **we propose a framework, DANCE, which improves the robustness of saliency methods** by following a two-step procedure. **First, we introduce a perturbation mechanism** that subtly varies the input sample without changing its intermediate representations. Using this approach, we can gather a corpus of perturbed (“decoy”) data samples while ensuring that the perturbed and original input samples follow similar distributions. **Second, we compute saliency maps for the decoy samples and propose a new method to aggregate saliency maps.** With this design, we offset influence of gradient saturation. From a theoretical perspective, we show that the aggregated saliency map not only captures inter-feature dependence but, more importantly, is robust against previously described adversarial perturbation methods. Our empirical results suggest that, both qualitatively and quantitatively, DANCE outperforms existing methods in a variety of application domains.

Paper Survey

3) Search Keywords: Knockoff, Knockoff filter

Multilayer knockoff filter: Controlled variable selection at multiple resolutions [17]

Interpretable Classification of **Bacterial Raman Spectra** with Knockoff Wavelets [18]

Improving the Stability of the Knockoff Procedure: Multiple Simultaneous Knockoffs and Entropy Maximization [19]

Metropolized **Knockoff Sampling** [20]

Adjusting the Benjamini-Hochberg method for controlling the false discovery rate in knockoff assisted variable selection [21]

Knockoff filter-based feature selection for discrimination of non-small cell lung cancer **in CT image** [22]

Knockoff boosted tree for **model-free variable selection** [23]

Paper Survey

Output-Related and -Unrelated **Fault Monitoring** with an Improvement **Prototype Knockoff Filter** and Feature Selection Based on Laplacian Eigen Maps and Sparse Regression [24]

Compositional **knockoff filter for high-dimensional** regression analysis of microbiome data [25]

Normalizing Flows for Knockoff-free Controlled **Feature Selection** [26]

Power analysis of knockoff filters for correlated designs [27]

A **prototype knockoff filter** for **group selection** with FDR control [28]

Some **Analysis** of the **Knockoff Filter** and its **Variants** [29]

A **knockoff filter** for **high-dimensional** selective inference [30]

Panning for Gold: **Model-free Knockoffs** for **High-dimensional** Controlled Variable Selection [31]

Paper Survey

Controlling the FDR in variable selection via **multiple knockoffs** [32]

Searching for consistent associations with a **multi-environment** knockoff filter [33]

Sequential knockoffs for continuous and categorical predictors: with application to a **large Psoriatic Arthritis clinical trial pool** [34]

Bayesian Knockoff Filter Using Gibbs Sampler [35]

Financial factors selection with knockoffs: Fund replication, explanatory and prediction networks [36]

Novel learning framework (knockoff technique) to evaluate metric ranking algorithms to describe **human response to injury** [37]

Interpretable signal analysis with knockoffs enhances **classification of bacterial raman spectra** [38]

Paper Survey

Identification of **putative causal loci in whole-genome sequencing** data via knockoff statistics [39]

Controlling the False Discovery Rate for **Binary Feature Selection** via Knockoff [40]

A **pseudo knockoff filter** for correlated features [41]

False Discovery Rate Control in **Cancer Biomarker** Selection Using Knockoffs [42]

KIMI: Knockoff Inference for Motif Identification from **molecular sequences** with controlled false discovery rate [43]

The knockoff filter for FDR control in **group-sparse** and **multitask** regression [44]

Lasso, knockoff and Gaussian covariates: a **comparison** [45]

False Discovery Rate Controlled Heterogeneous **Treatment Effect** Detection for Online Controlled Experiments [46]

Paper Survey

GGM knockoff filter: False Discovery Rate Control for **Gaussian Graphical Models** [47]

A Pipeline for Variable Selection and False Discovery Rate Control With an Application in **Labor Economics** [48]

High-dimensional variable selection for **ordinal outcomes** with error control [49]

Discovering Conditionally **Salient Features** with Statistical Guarantees [50]

Paper Survey

About Generating Knockoffs: 5 papers

About Feature Selection (Similar to DeepPINK): 4 papers

Application Domain: 3 papers

My Interests: 3 papers

[8] Burns, Collin, Jesse Thomason, and Wesley Tansey. "Interpreting black box models via hypothesis testing." **Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference. 2020.**

[14] Yoon, Jinsung, James Jordon, and Mihaela van der Schaar. "INVASE: Instance-wise variable selection using neural networks." **International Conference on Learning Representations. 2018.**

[15] Tang, Yehui, et al. "Scop: Scientific control for reliable neural network pruning." arXiv preprint arXiv:2010.10732 (2020).

중견 연구 제안서 작성

- 국내 의료 AI 관련 국책 과제 조사

  - 보건복지부가 주관하여 2025년까지 향후 5년간 총 945억원 사용해, **의료데이터셋**과 **의료 인공지능 기술** 개발을 진행함. => 총 3가지 사업으로 요약됨.

1) **중환자 특화 빅데이터** (Korean Medical Information Mart for Intensive Care, K-MIMIC) 구축 및 AI 기반 임상 **의사결정 지원** 시스템 (Clinical Decision Support System, CDSS) 개발
   - 데이터 정제를 통해 데이터 **사용자 수요에 따라 제공 가능**한 데이터셋 구성
   - 확보된 중환자 데이터를 **공유 분석하는 플랫폼** 구축
   - 중환자 CDSS를 위한 **설명 가능한 AI 알고리즘** 개발: 임상적 인과관계에 근거하여 중증도 평가, 재입원 가능성 등을 설명할 수 있는 AI 알고리즘 개발
   - **임상 활용 검증 및 실증** 연구

중견 연구 제안서 작성

2) **디지털 병리** 기반의 **암 전문** AI 분석 솔루션 개발

: 병리 진단이 필수적인 암에 대하여, AI 병리 분석 솔루션 개발

=> 병리 판독 업무 효율화, 환자 진단 및 치료 효과성 제고

- 병리 데이터, **주석 처리**를 비롯한 **표준화 처리**, 의료 기관 간의 이미지 색상 차이, 스캐너 제조사에 따른 차이 등을 위한 표준화 기술 개발

- **설명 가능한 AI 알고리즘** 기반 SW 개발, **임상 검증 및 활용 평가**

3) **합성 의료 데이터** 생성 및 **임상 활용 실증** 연구

: 민감 정보인 의료 데이터 정보 보호 및 활용성 제고를 위해, **공개 목적의 합성 의료 데이터 생성** 및 임상적 유효성 및 활용성 검증 연구 지원

- 시나리오 기반 고품질 데이터셋 구축: **합성 데이터 생성 모델** 개발

- 합성 데이터의 **임상적 유효성 및 활용성 실증** 연구

중견 연구 제안서 작성

요약 및 의견:

현재 국가적 차원에서 관심 있어 하는 내용은,

- 실생활에서 사용 가능, 설명 가능한 의료 AI 개발 => 진단 효율성 증대 및 의료 서비스 격차를 줄임.

가 하나의 최종 목표로 보이고, 이를 위한 기반 사업으로,

- 데이터셋 표준화 기술 및 공유 플랫폼, 합성 의료 데이터 생성 기술 개발

등을 진행하고자 함을 알 수 있었고,

이와 같은, 국가적 차원의 큰 그림에서, 당장 표준화된 데이터셋을 확보하기는 어려울 수 있으나, 전처리 과정과 표준화 방식, 학습 모델 제안 등을 포함한 우리만의 프레임워크가, 실생활에서 어느 정도 사용 가능한 세그멘테이션 성능을 보여줌으로써, 추후 표준화 기술에 대한 방향성 제시나, 국가 표준 세그멘테이션 모형의 초석을 다지는 데에 이바지함이 우리의 역할로 있을 수 있음.

중견 연구 제안서 작성

- 가제:

  한글: 의료 분야내 주요 관심 구조물 세그멘테이션을 위한 사전훈련 최신 딥러닝 모델 개발

  영문: Development of Pre-trained Deep Learning Model for Segmentation of Structures of Major Interest in the Medical Field

- 연구 목표:

  본 과제에서는 몇몇 주요한 해부학적 구조에 대해서 세계 최고 수준의 성능을 갖는 사전학습된 세그멘테이션 모델을 배포함으로써, 다양한 의료분야 응용에 성공적으로 기여하는 것을 최종 목표로 하며, 이를 위해 다음과 같은 연차별 연구 목표를 설정하여 추진한다.

중견 연구 제안서 작성

1. 연구의 목표 (2페이지)

    - 최종 연구 목표

    - 연차별 목표 및 개요도

    - 연차별 연구 개발 내용, 창의성 및 도전성, 추진 전략, 주요 결과물

2. 연구의 필요성 (2페이지)

    - 4~5개로 나누어 총 2페이지

3. 연구자의 연구 수행역량 (1페이지)

    - 연구실 스펙 및 노하우 등 자랑거리

중견 연구 제안서 작성

4. 연구의 추진 전략 및 방법 (연차별 1.5페이지)

    - 연차별 상세 내용 및 주요 기술 설명 (그림 포함하여, 연차당 3가지 정도)

5. 연구결과의 중요성 (0.5페이지)

    - 학문적/기술적 파급 효과

    - 경제적/사회적 파급 효과

6. 연구기간 및 연구비 적정성 (0.5페이지)

    - 추진일정 등

    - 연구비 적정성

중견 연구 제안서 작성

데이터셋 이름	촬영 범위	촬영 방식	차원	타겟	샘플수
CAU Orbit	머리, 안와 부근	CT	3D	안와내 구조물	1237
SNU CMC Gallbladder	담낭	Ultrasound	2D	담낭 용종	537
HNSCC	머리와 목	CT	3D	두경부암	627 (추가 논의)
Prostate MRI US Biopsy	전립선 부근	MR	3D	전립선	1151 (추가 논의)
CBIS DDSM	유방	Mammography	2D	유방암**(예비)**	1566 (추가 논의)
TCGA KIRC	신장 부근	CT	3D	신장	267+ **(추가 계획)**
TCGA GBM	머리, 뇌	MR	3D	뇌종양	262+ **(추가 계획)**
LIDC IDRI	흉부, 폐	CT	3D	폐 결절	1010 (추가 논의)
ACRIN 6664	결장 부근	CT	3D	결장 암	825 (추가 논의)
ISIC 2017	피부	Dermoscopic image	2D	피부 병변	2600 (추가 논의)

중견 연구 제안서 작성

- 다음 진행 계획:

  - 각 조사된(할) 참고 논문 기반, 각 세그멘테이션 연구가 어떤 면에서 중요할지(어떤 사업성을 지닐지) 조금 더 구체적으로 알아보기

    -> 예를 들어, 유방암 자동 진단 연구는 꽤 있었으나, 세그멘테이션 자체는 중요하게 다루어지지 않고 있을 수 있음.

    -> Ref 없이는 해당 구조물 세그멘테이션의 중요성 설명이 어려움

  - 국책 사업과의 관련성 구체화를 통한 연구 필요성 극대화 (강점 살리기)

    -> 국가적으로 중요하게 다뤄지는 부분(연구, 부위, 질병 등)과 연관이 깊다면, 강조되어 서술될 수 있음.

중견 연구 제안서 작성

- 다음 진행 계획:

  - 샘플 수 보충 계획 구체화

    -> 서로 다른 출처의 데이터셋을 병합할 계획, 이에 따른 긍정적&부정적 효과 검토

    - 긍정적: 국책 사업에서도, 다수의 단체로부터 데이터셋 정제 및 표준화 진행할 계획임. -> 우리의 연구가 그에 대한 초석이 될 수 있음.

    - 부정적: 수반되는 부정적 효과들을 당장 가늠하기 어려움.

      -> 같은 CT 촬영 방식일지라도, 다양한 촬영기기, 인종 등에 따른 고려가 어려움.

      -> 데이터 전처리를 위한 시간 소모, 현재로썬 가늠하기 어려움.

    => 부정적 효과가 있음에도, 제안서의 메리트가 될 수 있음. 논의 필요.

갑상선 연구

- 새로운 AX 1, 2, 4, 5 세그멘테이션 작업

  -> 5,870 / 11,480 장 완료

산학 협력 프로젝트

- 특허 출원: 특허 담당자 분 메일 수신 대기중

Reference

[1] Romano, Yaniv, Matteo Sesia, and Emmanuel Candès. "Deep knockoffs." Journal of the American Statistical Association 115.532 (2020): 1861-1872.

[2] Sudarshan, Mukund, Wesley Tansey, and Rajesh Ranganath. "Deep direct likelihood knockoffs." Advances in neural information processing systems 33 (2020): 5036.

[3] Duarte, Marco F., and Siwei Feng. "Knockoff-Inspired Feature Selection via Generative Models." (2019).

[4] Zhao, Xuebin, et al. "Distribution-dependent feature selection for deep neural networks." Applied Intelligence (2021): 1-11.

[5] Huang, Yingkun, et al. "Supervised feature selection through Deep Neural Networks with pairwise connected structure." Knowledge-Based Systems 204 (2020): 106202.

[6] using a Teacher, Deep Feature Selection. "Deep-gKnock: nonlinear group-feature selection with deep neural network."

[7] Laksshmi, KS Gowri, and R. Umagandhi. "An Effective Pulmonary Nodule Detection In Ctand X-Ray Images Using Enhanced Inception-Residual Convolutional Neural Network." Annals of the Romanian Society for Cell Biology (2020): 133-148.

[8] Burns, Collin, Jesse Thomason, and Wesley Tansey. "Interpreting black box models via hypothesis testing." Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference. 2020.

Reference

[9] Jordon, James, Jinsung Yoon, and Mihaela van der Schaar. "KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks." International Conference on Learning Representations. 2018.

[10] Song, Zixuan, and Jun Li. "Variable selection with false discovery rate control in deep neural networks." Nature Machine Intelligence 3.5 (2021): 426-433.

[11] Fan, Zhenjiang, et al. "Causal inference using deep-learning variable selection identifies and incorporates direct and indirect causalities in complex biological systems." bioRxiv (2021).

[12] Guo, Juncheng, et al. "An embedded gene selection method using knockoffs optimizing neural network." BMC bioinformatics 21.1 (2020): 1-19.

[13] Liu, Ying, and Cheng Zheng. "Auto-encoding knockoff generator for FDR controlled variable selection." arXiv preprint arXiv:1809.10765 (2018).

[14] Yoon, Jinsung, James Jordon, and Mihaela van der Schaar. "INVASE: Instance-wise variable selection using neural networks." International Conference on Learning Representations. 2018.

[15] Tang, Yehui, et al. "Scop: Scientific control for reliable neural network pruning." arXiv preprint arXiv:2010.10732 (2020).

[16] Lu, Yang, et al. "Robust Decoy-enhanced Saliency Maps." arXiv preprint arXiv:2002.00526 (2020).

Reference

[17] Katsevich, Eugene, and Chiara Sabatti. "Multilayer knockoff filter: Controlled variable selection at multiple resolutions." The annals of applied statistics 13.1 (2019): 1.

[18] Chia, Charmaine, et al. "Interpretable Classification of Bacterial Raman Spectra with Knockoff Wavelets." IEEE Journal of Biomedical and Health Informatics (2021).

[19] Gimenez, Jaime Roquero, and James Zou. "Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization." The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019.

[20] Bates, Stephen, et al. "Metropolized knockoff sampling." Journal of the American Statistical Association (2020): 1-15.

[21] Sarkar, Sanat K., and Cheng Yong Tang. "Adjusting the Benjamini-Hochberg method for controlling the false discovery rate in knockoff assisted variable selection." arXiv preprint arXiv:2102.09080 (2021).

[22] Li, Xiaomei, et al. "Knockoff filter-based feature selection for discrimination of non-small cell lung cancer in CT image." IET Image Processing 13.3 (2019): 543-548.

[23] Jiang, Tao, Yuanyuan Li, and Alison A. Motsinger-Reif. "Knockoff boosted tree for model-free variable selection." Bioinformatics 37.7 (2021): 976-983.

Reference

[24] Xue, Cuiping, Tie Zhang, and Dong Xiao. "Output-Related and-Unrelated Fault Monitoring with an Improvement Prototype Knockoff Filter and Feature Selection Based on Laplacian Eigen Maps and Sparse Regression." ACS omega 6.16 (2021): 10828-10839.

[25] Srinivasan, Arun, Lingzhou Xue, and Xiang Zhan. "Compositional knockoff filter for high-dimensional regression analysis of microbiome data." Biometrics (2020).

[26] Hansen, Derek, Brian Manzo, and Jeffrey Regier. "Normalizing Flows for Knockoff-free Controlled Feature Selection." arXiv preprint arXiv:2106.01528 (2021).

[27] Liu, Jingbo, and Philippe Rigollet. "Power analysis of knockoff filters for correlated designs." arXiv preprint arXiv:1910.12428 (2019).

[28] Chen, Jiajie, Anthony Hou, and Thomas Y. Hou. "A prototype knockoff filter for group selection with FDR control." Information and Inference: A Journal of the IMA 9.2 (2020): 271-288.

[29] Chen, Jiajie, Anthony Hou, and Thomas Y. Hou. "Some analysis of the knockoff filter and its variants." arXiv preprint arXiv:1706.03400 (2017).

[30] Barber, Rina Foygel, and Emmanuel J. Candès. "A knockoff filter for high-dimensional selective inference." The Annals of Statistics 47.5 (2019): 2504-2537.

[31] Candès, Emmanuel Jean, et al. Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. Department of Statistics, Stanford University, 2016.

Reference

[32] Emery, Kristen, and Uri Keich. "Controlling the FDR in variable selection via multiple knockoffs." arXiv preprint arXiv:1911.09442 (2019).

[33] Li, Shuangning, et al. "Searching for consistent associations with a multi-environment knockoff filter." arXiv preprint arXiv:2106.04118 (2021).

[34] Kormaksson, Matthias, et al. "Sequential knockoffs for continuous and categorical predictors: With application to a large psoriatic arthritis clinical trial pool." Statistics in Medicine 40.14 (2021): 3313-3328.

[35] Gu, Jiaqi, and Guosheng Yin. "Bayesian Knockoff Filter Using Gibbs Sampler." arXiv preprint arXiv:2102.05223 (2021).

[36] Challet, Damien, Christian Bongiorno, and Guillaume Pelletier. "Financial factors selection with knockoffs: fund replication, explanatory and prediction networks." Physica A: Statistical Mechanics and its Applications (2021): 126105.

[37] Banerjee, Anjishnu, et al. "Novel learning framework (knockoff technique) to evaluate metric ranking algorithms to describe human re

[38] Chia, Charmaine, et al. "Interpretable signal analysis with knockoffs enhances classification of bacterial raman spectra." arXiv preprint arXiv:2006.04937 (2020).

[39] He, Zihuai, et al. "Identification of putative causal loci in whole-genome sequencing data via knockoff statistics." Nature communications 12.1 (2021): 1-18.

Reference

[40] Xie, Yuxiang, and Kwun Chuen Gary Chan. "Controlling the False Discovery Rate for Binary Feature Selection via Knockoff." arXiv preprint arXiv:1807.00931 (2018).

[41] Chen, Jiajie, Anthony Hou, and Thomas Y. Hou. "A pseudo knockoff filter for correlated features." Information and Inference: A Journal of the IMA 8.2 (2019): 313-341.

[42] Shen, Arlina, et al. "False discovery rate control in cancer biomarker selection using knockoffs." Cancers 11.6 (2019): 744.

[43] Bai, Xin, et al. "KIMI: Knockoff Inference for Motif Identification from molecular sequences with controlled false discovery rate." Bioinformatics 37.6 (2021): 759-766.

[44] Dai, Ran, and Rina Barber. "The knockoff filter for FDR control in group-sparse and multitask regression." International conference on machine learning. PMLR, 2016.

[45] Davies, Laurie. "Lasso, knockoff and Gaussian covariates: a comparison." arXiv preprint arXiv:1805.01862 (2018).

[46] Xie, Yuxiang, Nanyu Chen, and Xiaolin Shi. "False discovery rate controlled heterogeneous treatment effect detection for online controlled experiments." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.

[47] Li, Jinzhou, and Marloes H. Maathuis. "GGM knockoff filter: False Discovery Rate Control for Gaussian Graphical Models." arXiv preprint arXiv:1908.11611 (2019).

Reference

[48] Klose, Sophie-Charlotte, and Johannes Lederer. "A Pipeline for Variable Selection and False Discovery Rate Control With an Application in Labor Economics." arXiv preprint arXiv:2006.12296 (2020).

[49] Fu, Han, and Kellie J. Archer. "High-dimensional variable selection for ordinal outcomes with error control." Briefings in bioinformatics 22.1 (2021): 334-345.

[50] Gimenez, Jaime Roquero, and James Zou. "Discovering conditionally salient features with statistical guarantees." International Conference on Machine Learning. PMLR, 2019.